PDB Assignment 2 Exercise 2.1

2025-11-14

Name1: Adwitiya Argha Priyadarshini Boruah
Matriculation No. 1: 7070291
Email1: adbo00002@stud.uni-saarland.de

Name2: Md Mobashir Rahman
Matriculation No. 2: 7059086
Email2: mdra00001@stud.uni-saarland.de

Exercise 2.1: DNA methylation in hematopoieses and clustering

a.

The file "methylation.csv" contains average DNA methylation levels for many genomic regions across different multiple blood and skin cell types. Before analysis, several formatting must be corrected. Methylation values are stored with commas instead of decimal points, and missing entries are represented by ".", which prevents direct numerical conversion. All the methylation columns (columns 7-26) are therefore processed by replacing "." with missing values, converting commas to decimal points, and transforming the netries into numeric format. Missing methylation values are then set to 0, as required in the assignment, THe first six cilumns containing genomic annotations remain unchanged. This results in a fully numeric methylation matrix suitable for computing distances and performing heirarchical clustering in the subsequent tasks.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(knitr)

#Reading the raw methylation file
#the file uses semicolons instead of commas
raw_data <- read_delim("methylation.csv", delim = ";", col_types = cols(.default = "c"))

#Identifying the methylation colums
#columns 1-6 are annotation, column 7 onwards it is numeric
methylation_columns <- colnames(raw_data)[7:ncol(raw_data)]
```

```r
#making a copy of the data for cleaning
clean_data <- raw_data

#Cleaning methylation columns:
# -replacing "." with NA (missing), replacing "," with "." (decimal), converting to numeric
clean_data[methylation_columns] <- clean_data[methylation_columns] %>%
  mutate(across(everything(), ~ .x
                %>% na_if(".") %>% #"." -> NA
                  str_replace(",", ".") %>% #"0,842" -> "0.842"
                  as.numeric())) #converting to numeric

#replacing the missing values in methylation columns with 0.
clean_data[methylation_columns][is.na(clean_data[methylation_columns])] <- 0
write_csv(clean_data, "methylation_clean.csv")

clean_data %>%
    select(chrom, chromStart, chromEnd, geneName, HSC, MPP1, MPP2) %>% head(10) %>% kable()
```

chrom	chromStart	chromEnd	geneName	HSC	MPP1	MPP2
chr1	3027000	3028000	NA	0.815	0.840	0.831
chr1	3140000	3141000	NA	0.822	0.831	0.897
chr1	3266000	3267000	NA	0.941	0.954	0.953
chr1	3291000	3292000	NA	0.889	0.881	0.881
chr1	3334000	3335000	NA	0.808	0.829	0.828
chr1	3612000	3613000	NA	0.935	0.937	0.943
chr1	3660000	3661000	NA	0.031	0.018	0.021
chr1	3661000	3662000	Xkr4	0.020	0.014	0.010
chr1	3681000	3682000	NA	0.809	0.837	0.844
chr1	3835000	3836000	NA	0.963	0.936	0.928

b.

Average methylation levels for each cell tpe were obtained by taking the mean across all genomic regions in the cleaned dataset. All annotation columns were excluded, adn only methylation columns (columns 7-26) were summarised. The resulting values represent the global methylation of each cell type . The computed averages were stroed for later analysis.

```r
#now loading the cleaned dataset
clean_data <- read_csv("methylation_clean.csv")
```

```
## Rows: 95086 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): chrom, geneName, ensemblId
## dbl (23): chromStart, chromEnd, name, cpgMinCoverage, HSC, MPP1, MPP2, CLP, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#identifying the methylation columsn again
methylation_columns_clean <- colnames(clean_data)[8:ncol(clean_data)]

mean_methylation <- clean_data %>%
  summarise(across(all_of(methylation_columns_clean), mean))

t(mean_methylation)
```

```
##             [,1]
## HSC    0.5961406
## MPP1   0.6170356
## MPP2   0.6288504
## CLP    0.6402188
## CMP    0.6417423
## GMP    0.6350850
## MEP    0.6070041
## CD4    0.6469553
## CD8    0.6480897
## B_cell 0.6460829
## Eryth  0.5164726
## Granu  0.6211135
## Mono   0.6259272
## TBSC   0.5682913
## ABSC   0.5694599
## MTAC   0.5382875
## CLDC   0.5465432
## EPro   0.5850959
## EDif   0.5822801
```

```r
write_csv(mean_methylation, "mean_methylation.csv")
```

Global methylation averages were calculated for each cell type to compare epigenetic patterns across hematopoietic and skin lineages. Early blood progenitors (HSC: 0.596, MPP1: 0.617, MPP2: 0.629) showed relatively low methylation, whereas differentiated immune cell types such as CLP (0.640), CMP (0.642), GMP (0.636), CD4 (0.647), CD8 (0.648), B-cell (0.646), granulocytes (0.621) and monocytes (0.626) displayed higher methylation levels. This trend reflects the general principle that methylation increases during lineage commitment as developmental potential decreases.

The erythroid lineage (EPro: 0.586, EDif: 0.582, Eryth: 0.516) deviated from this pattern by exhibiting notably lower methylation levels. This behaviour is characteristic of erythroid differentiation, which involves lineage-specific demethylation and activation of erythroid regulatory programs.

Skin-derived cell types (TBSC: 0.569, ABSC: 0.569, MTAC: 0.538, CLDC: 0.547) formed a distinct group with relatively high and internally consistent methylation levels, clearly separating them from the hematopoietic lineage. These observations agree with the developmental structure shown in Figure 1 and indicate that average methylation levels capture major differences between cell lineages, even though specific lineages do not follow a perfectly monotonic trend.

Different genomic regions follow characteristic methylation patterns. Promoter CpG islands of active genes are usually unmethylated, while promoters of genes that are switched off become methylated during differentiation. Active enhancers also tend to be hypomethylated, whereas inactive enhancers show higher methylation levels. Gene bodies typically carry moderate methylation, which can be associated with ongoing transcription. In contrast, intergenic regions are generally highly methylated to prevent unwanted transcription. These regional patterns help explain why global methylation increases as cells become more specialized: more promoters and enhancers become permanently silenced during lineage commitment.

c.1.

To compare methylation patterns between cell types, the methylation matrix was first constructed by selecting all methylation columns (one column per cell type). Each column therefore represents the methylation profile of that cell type across all genomic regions. Euclidean distances between all pairs of cell types were then computed. Since dist() calculates distances between rows, the matrix was transposed so that each column that is the cell type was treated as a sample.

```
#identifying methylation column i.e. the cell types
cell_type_columns <- colnames(clean_data)[8:ncol(clean_data)]

#building a numeric matrix of methylation values
#rows are genomic regions, columns are cell types
methylation_matrix <- clean_data %>%
  select(all_of(cell_type_columns))  %>%
  as.matrix()

#computing euclidean distance between cell types
#transpose is required because dist() computes row distances
distance_matrix <- stats::dist(t(methylation_matrix), method = "euclidean")

#converting to matrix for easier indexing
distance_matrix <- as.matrix(distance_matrix)
pairwise_distances <- distance_matrix[c("HSC", "CD8", "MTAC"), c("HSC", "CD8", "MTAC")]
pairwise_distances
```

```
##           HSC      CD8     MTAC
## HSC   0.00000 46.06283 78.55019
## CD8  46.06283  0.00000 81.77717
## MTAC 78.55019 81.77717  0.00000
```

This produces a full pairwise distance matrix, where small values indicate similar methylation patterns and large values indicate dissimilar cell types. HSC and CD8 show a moderate distance (46.06), indicating similarity within the blood lineage. MTAC (skin lineage) is much further from both HSC (78.55) and CD8 (81.78),which reflects clear separation between blood and skin methylation profiles.

c.2.

To merge clusters during hierarchical clustering, the average linkage criterion was implemented.

```
#average linkage between the two clusters A and B
#A and B are character vectors containing cell type names
average_linkage <- function(A, B, distance_matrix){

  #extracting distances for all combinations a € A, b € B
  pairwise_distances <- distance_matrix[A, B]

  #computing the average of these distances
  mean(pairwise_distances)
} #A small value of L(A,B) means the two clusters are very similar.
```

This function enables the clustering algorithm to evaluate the similarity between any two sets of cell types and choose the pair that should merge next.

c.3.

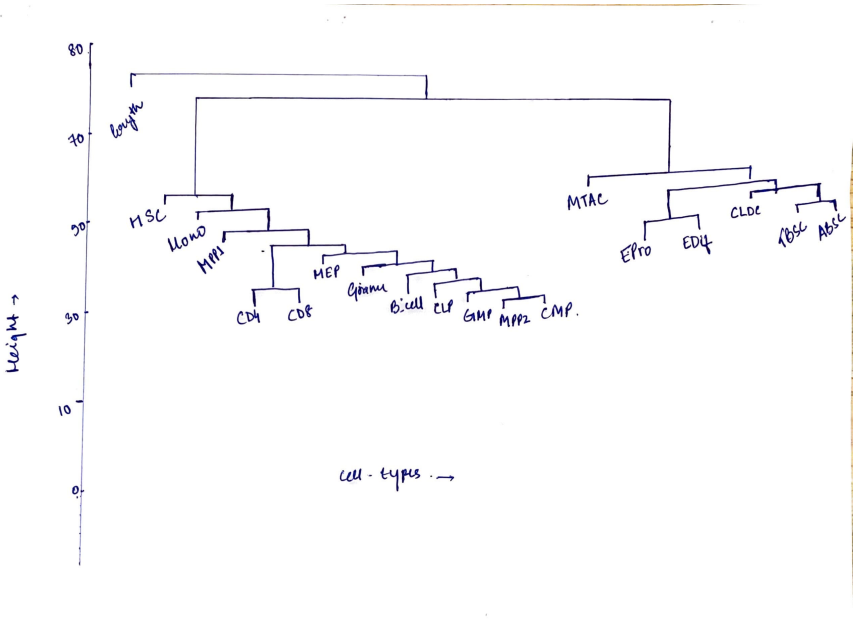
Each cell type was initialized as its own cluster. At every iteration, all currently existing clusters were compared, and the pair with the smallest average linkage value was merged. This procedure was repeated until only one cluster remained. For every merge, the algorithm printed:- * the two clusters being merged * their linkage value * the updated clustering structure

```
#initialising clusters: each cell types is its own cluster which is a list with 1 element
clusters <- as.list(cell_type_columns)
names(clusters) <- cell_type_columns

merge_history <- list() #to save the sequence of merges

while (length(clusters) > 1) {
  cluster_names <-  names(clusters)
  best_A <- NULL
  best_B <- NULL
  best_L <- Inf

  #comparing all paired of clusters
  for (i in 1:(length(clusters) - 1)) {
    for (j in(i + 1):length(clusters)) {
      A <- clusters[[i]]
      B <- clusters[[j]]

      #now computing average linkage
      Linkage_average_value <-  average_linkage(A, B, distance_matrix)
      #keeping track of smallest linkage
      if (Linkage_average_value < best_L) {
        best_L <- Linkage_average_value
        best_A <- cluster_names[i]
        best_B <- cluster_names[j]
      }
    }
  }

  #printing result of this merge
  cat("Merging:", best_A,"and", best_B, "with L =", best_L, "\n")

  #saving merge
  merge_history <- append(merge_history, list(list(A = best_A, B = best_B, L = best_L)))

  #merging the clusters
  new_cluster <- c(clusters[[best_A]], clusters[[best_B]])
  clusters[[best_A]] <- new_cluster
  clusters[[best_B]] <- NULL
}
```

```
## Merging: MPP2 and CMP with L = 15.64175
## Merging: CD4 and CD8 with L = 16.09429
## Merging: MPP2 and GMP with L = 19.60319
## Merging: MPP2 and CLP with L = 22.20841
## Merging: MPP2 and B_cell with L = 23.97957
## Merging: MPP2 and Granu with L = 25.23867
## Merging: EPro and EDif with L = 25.62158
## Merging: MPP2 and MEP with L = 26.45399
```

```
## Merging: MPP2 and CD4 with L = 28.34874
## Merging: MPP1 and MPP2 with L = 31.34756
## Merging: TBSC and ABSC with L = 33.76158
## Merging: MPP1 and Mono with L = 36.51274
## Merging: TBSC and CLDC with L = 37.88632
## Merging: TBSC and EPro with L = 38.61244
## Merging: HSC and MPP1 with L = 41.67949
## Merging: TBSC and MTAC with L = 44.26883
## Merging: HSC and TBSC with L = 71.58857
## Merging: HSC and Eryth with L = 76.52372
```

```
knitr::include_graphics("dendogram.png")#the hand draw dendrogram is shown below
```



Biological interpretation-

The hierarchical clustering clearly separates the hematopoietic i.e. blood and skin-derived lineages.

In the dendrogram, the skin cell types (TBSC, ABSC, MTAC, CLDC) form a distinct cluster that branches off independently from all blood-related cell types. This shows that global methylation profiles are sufficiently different to separate the two developmental systems. Within the hematopoietic branch, the clustering partially reflects the expected developmental succession. Early progenitors (HSC, MPP1, MPP2) remain closer to each other and merge before the committed immune lineages (CLP, CMP, GMP, B-cell, CD4, CD8, Granu, Mono). This agrees with the biological hierarchy, where multipotent progenitors differentiate into more specialized immune cells.

However, the erythroid lineage (Eryth, EPro, EDif) does not follow the expected order from Figure 1. These cells cluster separately and at an early height due to their unusually low global methylation, which differs from the increasing-methylation trend seen in other blood cell types.